マスク着用環境下での音声と画像を利用した発話推定

稲留　浩太郎[1]　嶋田　和孝[2]

概要：発話推定とは，VAD(Voice Activity Detection) と呼ばれる対象の話者が発話しているか否かの推定を行うタスクである．VAD には，音声情報だけでなく画像情報も利用して発話推定を行う AV-VAD(Audio-Visual VAD) が存在する．AV-VAD の多くは，画像の特徴量として口唇の部分を特徴量として使用している．そのため，マスクを着用した環境下においては従来の AV-VAD の手法が利用できない．本研究では，マスク着用環境下でも利用可能な AV-VAD として，オプティカルフローを利用して取得したマスクの動きを画像の特徴量として利用する AV-VAD の手法を提案する．

キーワード：機械学習，画像分類，話者・言語識別

Voice Activity Detection Under Mask-Wearing Conditions Using Audio and Image Information

Abstract: Voice Activity Detection (VAD) is a task to detect whether a target speaker is speaking or not. VADs include AV-VAD (Audio-Visual VAD), which use audio and image information to detect voice activity. Most AV-VADs use the lips as image features. Therefore, conventional AV-VAD methods cannot be used under mask-wearing conditions. In this study, we propose an AV-VAD method that can be used under mask-wearing conditions, which uses the mask motion acquired using optical flow as an image feature.

Keywords: Machine Learning,Image Classification,Speaker and Language Identification

1. はじめに

発話推定とは，VAD(Voice Activity Detection) と呼ばれる対象の話者が発話しているか否かの推定を行うタスクである．発話推定は主に対話システムや会話の分析等で利用される．音声情報を解析し，発話推定を行う研究は多くの手法で行われてきた [1]．昨今においても，Ali ら [2] による教師なし学習による手法や Alimi ら [3] による低コストな手法，Prithvi ら [4] による動的学習を使用した手法，といったように多くの研究が行われている．これらのような音声ベースの発話推定は A-VAD(Audio-based Voice Activity Detection) と呼ばれている．

[1] 九州工業大学 大学院情報工学府
Department of Creative Informatics, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
[2] 九州工業大学 大学院情報工学研究院 知能情報工学研究系
Department of Artificial Intelligence, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN

発話推定における課題の 1 つとして，環境音が存在する状況での発話推定を考える必要がある．環境音が存在する場合，発話の音声に環境音が混じりノイズが混入してしまうためである．発話推定の手法として，Hengshun ら [5] や Fei ら [6] のように画像の情報も利用して発話推定を行う AV-VAD(Audio-Visual Voice Activity Detection) も存在する．AV-VAD は，画像の特徴量を加えた発話推定を行うことで，ノイズが存在する環境下においても高い精度での推定を行うための手法である．発話推定は A-VAD，AV-VAD ともに今なお研究が行われている重要なタスクである．

昨今においては，新型コロナウイルスの影響もあり，今なおマスクを着用しなければならない場面が多くなっている．発話推定を行っている場面においてもマスクを着用した場合の影響を考える必要がある．元吉ら [7] や二宮ら [8] のような口唇を抽出して発話推定を行う AV-VAD では，マスクを着用した場合口唇がマスクで覆われるため，推定が困難になる．そのため，マスク着用環境下においても利用

可能な新たな AV-VAD が必要である．

マスク着用環境下での AV-VAD を行うにあたり，画像の特徴量として何を利用するか考える必要がある．マスクを着用した環境下で発話を行った場合，マスクが動き陰影の変化が生じる．このマスクの動きを特徴量として利用することにより，環境音などのノイズに耐性を持った発話推定ができる可能性がある．そこで本研究では，オプティカルフローを利用して取得したマスクの動きを画像の特徴量として，また音声データの周波数ごとの音の大きさを音の特徴量として利用するマスク着用環境下においても利用可能な AV-VAD の手法を提案する．

2. データセット

本節では，マスク着用環境下の発話推定を行うために作成したデータセットの説明を行う．2.1 項では撮影の条件や，撮影環境についての説明を行う．また，2.2 項では撮影した動画の内容についての説明を行う．

2.1 データセットの内容

本項ではマスク着用環境下における発話推定のために作成したデータセットの撮影環境についての説明を行う．本データセットは 14 名の大学生及び大学院生が各々の PC に付属する web カメラを用いて自身が発話する様子を撮影したものである．14 名の学生のうち，13 名は図 1 のように白色の不織布マスクを着用して撮影を行う．また，1 名は図 2 のように白色のウレタンマスクを着用した状態で撮影を行う．撮影した動画は，不織布マスクを着用して環境音の無い静かな環境で撮影を行った動画が 13 個，不織布マスクを着用して環境音がありノイズが存在する環境で撮影した動画が 10 個，ウレタンマスクを着用して環境音が無い静かな環境で撮影を行った動画が 1 個，ウレタンマスクを着用して環境音がありノイズが存在する環境で撮影した動画が 1 個，合計 25 個の動画である．撮影の際に使用する環境音は，撮影者が自由に選択したものである．本データセットの動画は学生が各々の PC で撮影を行ったため，フレームレートやサンプリングレートが異なる．本データセットの動画のフレームレートは 25 または 30 である．また，サンプリングレートは 32kHz，44.1kHz，48kHz，96kHz のいずれかである．

2.2 撮影した動画の内容

本項では撮影した動画の内容について説明を行う．撮影した動画の内容の概要を図 3 に示す．本データセットの動画はマスクを着用した状態で一定区間ごとに発話を行うといった内容である．本データセットの撮影時に発話する内容は，台本に沿ったものである．この台本は，動画内の発話区間を明確にするため，一定時間ごとに文章を読み上げるように構成されている．

図 1 不織布マスクを着用した動画の例

図 2 ウレタンマスクを着用した動画の例

読み上げる文章は 25 種類の文となる．表 1 に読み上げる文章の例を示す．1 文目から 15 文目は共通の文章を読み上げる．16 文目から 20 文目は読み上げる人により回答がことなる文章を読み上げる．この 1 文目から 20 文目の区間は図 3 の「1:環境音無し」にあたる．21 文目から 25 文目は共通の文章ではあるが，読み上げてもらった後に再び発声せずに発話の動きだけを行う．この区間は図 3 の「2:環境音無し・発話無し含」に該当する．また，21 文目から 25 文目は，環境音を流しながらもう一度撮影する．この区間は図 3 の「3:環境音あり・発話無し含」に該当する．以上が読み上げる文や条件に関する説明となる．

撮影した動画は台本の文章を読み上げた部分を発話，それ以外の部分を非発話として動画の各フレームにアノテーションを行う．アノテーションを行ったデータの内訳を表 2 に示す．また，データセット内のデータは環境音が無く発話を行わない部分を含まない区間，環境音が無く発話を行わない部分を含んだ区間，環境音がある区間の 3 種類のデータに分ける．区間ごとに分けたデータは，環境音がなく通常通りに発話する区間を「環境音無」，環境音がなく発話しない場合がある区間を「環境音無/発話無含」，環境音があり発話しない場合がある区間を「環境音有/発話無含」と定義する．

3. 提案手法

本節では実験で使用した手法についての説明を行う．

提案手法の概要を図 4 に示す．画像の特徴量にはマスクの動きとして，オプティカルフローを使用する．マスクの領域は顔検出を行い，取得した顔の領域をもとに推定したマスクの領域を使用する．その後，マスクの動きとして，

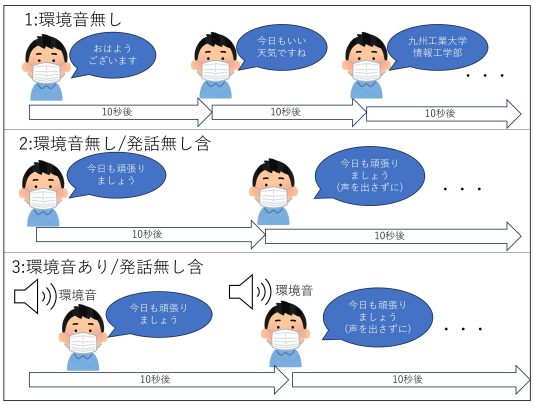

図 3 データセット撮影の概要

表 1 台本の内容例

1 文目	おはようございます
2 文目	今日もいい天気ですね
3 文目	九州工業大学情報工学部
4 文目	今は 12 時過ぎです
5 文目	よろしくお願いします
16 文目	質問：自分の名前
17 文目	質問：自分の出身地
18 文目	質問：自分の出身高校名
19 文目	質問：好きな教科
20 文目	質問：今日食べた昼ごはん

表 2 アノテーションデータ数

環境音の有無	発話タグ属性	データ件数
環境音無し	発話	17360
環境音無し	非発話	114434
環境音あり	発話	3118
環境音あり	非発話	32067

推定したマスクの領域内のオプティカルフローを取得する．音声の特徴量には，周波数ごとの音の大きさを使用する．最後に，取得した特徴量を用いて LSTM に学習させ，発話推定を行う．

本節の 3.1 項，3.2 項では，画像の特徴量取得の際に使用する顔検出システムとオプティカルフローについての説明を行う．3.3 項では音声の特徴量の取得の際に使用する技術の説明を行う．3.4 項では，本研究で使用する LSTM についての説明を行う．また，3.5 項および 3.6 項では，本実験で使用する特徴量を取得する提案手法についての説明を行う．

3.1 顔検出

本研究ではマスクの領域の推定のために顔検出を行う．学習済みモデルである face-mask-detection-tf2[*1]を使用する．face-mask-detection-tf2 には学習済みモデルが用意されており，マスクを着用している状態においても顔検出が可能である．

[*1] https://github.com/PureHing/face-mask-detection-tf2

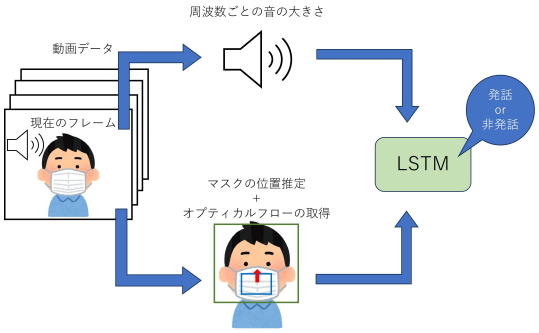

図 4 データセット撮影の概要

3.2 オプティカルフロー

本研究では，推定で使用する特徴量としてオプティカルフローを用いる．オプティカルフローとは，動画の連続する 2 フレーム間を比較することで，動画内の物体の動きを取得し，ベクトルで表現する技術である．本研究では発話時に生じるマスクの動きを特徴量として利用する．そのため，物体の動きをベクトルとして取得できるオプティカルフローを使用する．

オプティカルフローの計算は連続する 2 フレーム間においてピクセルの強度が一定という仮定を置いて行われる．例として，t フレーム目の座標 (x,y) にある強度 I のピクセルでの場合を考える．強度 I のピクセルが次のフレームの t+dt フレーム目で座標 (x+dx,y+dy) に移動したとする．この際ピクセルの強度 I は変化しないと仮定したため，次式

$$I(x, y, t) = I(x + dx, y + dy, t + dt)$$

が成立する．右辺をテイラー展開したのちに共通項を除き，dt で割ることにより

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0$$

が成立する．この u,v がオプティカルフローベクトル (u,v) となる．しかし，この式では 2 つの未知数に対して 1 つの方程式しか与えられていないため，ベクトル (u,v) の値を求めることが不可能である．このベクトル (u,v) を求める手段として，Bruce D.Lucas ら [9] が提唱した Lucas-Kanade 法や Gunnar Farneback[10] が提唱した Gunnar Farneback アルゴリズムが存在する．

本研究では OpenCV[*2]を使用してオプティカルフローの計算を行う．OpenCV では Lucas-Kanade 法によるオプティカルフローの取得及び追跡や，Gunnar Farneback アルゴリズムによる高密度オプティカルフローの取得が可能である．本研究においては，画像内の全ピクセルに対してオプティカルフローの計算をする高密度オプティカルフローが可能な Gunnar Farneback アルゴリズムを用いて特徴量を取得する．

[*2] https://docs.opencv.org/3.4/dc/d6b/group__video__track.html

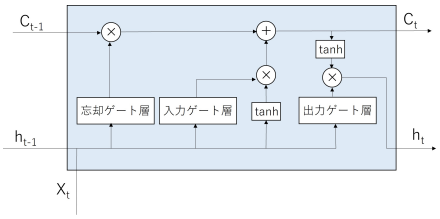

図 5 LSTM の概要図

3.3 音声情報

本研究では，動画内の音声情報も特徴量として使用する．音声情報の取得には Pydub[*3]を用いる．Pydub はオーディオファイルのサンプリングレートや raw audio data の取得だけでなく，サンプリングレートの変更も可能な，音声解析用のライブラリである．本研究の手法では，Pydub を用いて raw audio data を取得し，フーリエ変換を行うことで周波数ごとの音の大きさを特徴量として取得する．本研究においては，画像の情報と音声の情報両方扱うことで，互いのノイズの影響を抑えられると考える．そのため，[2] のような V-VAD で使用されているノイズに強い特徴量ではなく，フーリエ変換を行い周波数ごとの大きさを取得する手法を使用する．

3.4 LSTM

本研究では分類器として LSTM(Long Short Term Memory) を使用する．LSTM とは Hochreiter ら [11] の提唱を発端に発展した，ニューラルネットワークの一種である．LSTM の特徴は過去の情報を記憶しておくセルが存在する点である．LSTM は図 5 のように中間層が tanh 層，忘却ゲート層，入力ゲート層，出力ゲート層により構成される．LSTM では最初に忘却ゲート層で過去のセル C_{t-1} をどの程度保持するかの計算が行われる．次に，入力ゲート層と tanh 層により現在の入力を基に現在のセル C_t への更新を行う．最後に出力ゲート層と tanh 層により出力 h_t の計算が行われる．

発話推定を行うに当たり，現在のフレームの情報だけでなく，時系列データを使用しての推定が必要である．動画データを扱う機械学習を行う場合，長い時系列データを扱える LSTM が適している．そのため，本研究では Keras[*4]を使用して LSTM の構築を行う．

3.5 画像ベースの特徴量の取得

本項ではマスクのオプティカルフローを取得する手法について説明する．マスクのオプティカルフローを取得するには，マスクの領域を推定する必要がある．マスクの領域の推定は，顔検出を行い取得した顔の領域をもとに行う．顔検出には 3.1 項で説明した face-mask-detection-tf2 を使用する．顔検出と推定したマスクの領域の例を図 6 に示す．緑線で囲んだ範囲が顔検出で取得した顔の領域である．青線で囲んだ領域が取得した顔の領域から推定したマスクの領域である．

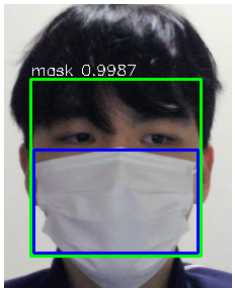

図 6 マスクの領域の推定

オプティカルフローの計算には 3.2 項で述べた OpenCV での高密度オプティカルフローを使用する．高密度オプティカルフローの計算により，1 ピクセルごとにベクトルの大きさと角度が取得できる．LSTM で使用する特徴量の次元数を統一するために，マスクの領域を縦横ごとに 10 分割を行い，100 個の領域に分け，各領域ごとのオプティカルフローの平均値を計算を行う．本研究では計算した各領域のオプティカルフローの平均値をマスクのオプティカルフローとして使用する．取得したマスクのオプティカルフローは，200 次元の特徴量となる．

3.6 音声ベースの特徴量の取得

本項では音声の周波数と大きさを取得する手法について説明する．動画の音声情報の取得には 3.3 項で説明した Pydub を使用する．本研究のデータセットはサンプリングレートの統一がされていないため，Pydub で音声情報を取得する際にサンプリングレートを 48kHz へ変換させる．解析を行う音声情報の区間は，現在のフレームから 0.04 秒前までの区間とする．指定した区間の音声情報にフーリエ変換を適用することで周波数ごとの音の大きさが得られる．本実験の場合，0.04 秒の区間をフーリエ変換しているため，25Hz ごとに大きさが得られる．本実験では一般的な声の範囲である 200Hz から 1000Hz の範囲内の音の大きさを特徴量として使用する．25Hz ごとに音の大きさが得られるため，本実験で取得できる音声ベースの特徴量は 33 次元の特徴量となる．

4. 実験

本節では，本研究で行った実験とその結果について説明を行う．4.1 項では実験の概要の説明を行う．4.2 項では，本研究で行った不織布マスクを着用した環境下での発話推定の結果について述べる．4.3 項では，追加実験として行った，ウレタンマスクを着用した環境下のデータに対しての発話推定の説明と追加実験の結果について述べる．

*3 https://github.com/jiaaro/pydub
*4 https://keras.io/api/

表 3 被験者ごとの訓練データ数とテストデータ数

被験者	訓練データ	テストデータ		
		環境音無	環境音無 発話無含	環境音有 発話無含
被験者 A	9708	2988	716	646
被験者 B	10252	1878	498	506
被験者 C	10328	2294	502	388
被験者 D	10248	2788	500	424
被験者 E	10496	1598	458	384
被験者 F	9944	2830	614	660
被験者 G	10204	2696	586	446
被験者 H	10064	1990	654	494
被験者 I	9844	2994	816	588
被験者 J	10288	1862	452	488

4.1 実験の概要

本項では本研究で行う実験の説明を行う．本実験は LSTM を使用し，3.5 項と 3.6 項で説明した特徴量を用いて発話推定を行う．本実験で使用した LSTM は 3.4 項で述べた Keras の LSTM を使用する．LSTM の設定はバッチサイズを 2048，エポック数を 100，学習率を 0.001，隠れ層のユニット数を 128 とする．また，シーケンス長は 4 として，現在フレームから 3 フレーム前までの 4 つの時点での 3.5 項と 3.6 項で説明した特徴量を使用する．以上の設定で，現在フレームが発話か非発話かの推定を行う．

本実験では有効性の確認のために，画像ベースの特徴量のみでの推定，音声ベースの特徴量のみでの推定，両方の特徴量を用いた推定の 3 通りの実験を行う．訓練データとテストデータには不織布のマスクを着用して撮影を行ったデータを使用する．テストデータは環境音がある環境下で撮影できた 10 人から 1 人を選択し，選択した人が撮影した動画のデータを使用する．訓練データには不織布マスクを着用した環境下のデータのうち，テストデータに使用しなかったデータを使用する．実験はテストデータで選択できる 10 人すべての場合で行う．

訓練データは，アノテーションを行ったデータから環境音がないデータと環境音があるデータで非発話のデータと発話のデータが同数になるようにランダムサンプリングを行う．また，テストデータはアノテーションを行ったデータから「環境音無」,「環境音無/発話無含」,「環境音有/発話無含」の各区間ごとに非発話のデータと発話のデータが同数になるようにランダムサンプリングを行う．表 3 に被験者ごとの訓練データ数とテストデータ数を示す．

4.2 結果

本項では使用した特徴量ごとに発話推定の結果の確認を行う．

4.2.1 画像ベースの特徴量のみでの推定

画像ベースの特徴量を使用し，「環境音無」のテストデータの推定結果を表 4,「環境音無/発話無含」のテストデータの推定結果を表 5,「環境音有/発話無含」のテストデータの推定結果を表 6 に示す．平均の値を見ると，F 値が一番大きい場合でも表 6 の 0.641 となっており，低い精度の推定となった．

表 5 の平均の値を見ると，表 4 の平均と比較して低い精度となっている．この理由は非発話のデータに口を動かすが発話をしない部分のデータが含まれており，音の情報がないため発話と非発話の区別がつかないため誤った推定をしたためだと考えられる．

表 6 は口を動かすが発話をしない場合が含まれているにも関わらず表 4 と比較して精度の低下が見られなかった．この理由は環境音が流れているため，音声が聞こえるようにはっきりと発音を行った結果，口の動きが大きくなり，連動してマスクの動きも大きくなったため発話の検知がしやすくなったと考えられる．

4.2.2 音声ベースの特徴量のみでの推定

音声ベースの特徴量を使用し,「環境音無」のテストデータを推定した結果を表 7,「環境音無/発話無含」のテストデータを推定した結果を表 8,「環境音有/発話無含」のテストデータを推定した結果を表 9 に示す．平均の値を見ると，どのデータも画像のみの推定と比べて高い精度となった．

表 8 では，表 7 と比較して精度の低下は見られなかった．これは，音声情報のみを使用しているため，口を動かすが発話をしない部分の影響が無かったためだと考えられる．

表 9 は他のデータ比べて精度が低くなっている．音声のみを用いた手法では，環境音と発話の区別ができず，精度が低下したと考えられる．

各被験者のデータを見た場合，表 9 において被験者 C，G，H，I が環境音が無い環境である表 7 と表 8 の結果と比較して大きく精度が下がっている．表 9 の被験者 C，G，H，I においては precision が低く recall が高くなっており，発話の誤検知が多かったことがわかる．被験者 C，G，H，I は環境音として歌や，演奏を大きな音で流しており，200Hz から 1000Hz 内の音が鳴り，発話と誤検知したと考えられる．精度が低下しなかった被験者では，環境音として歌が流れているものもあったが小さい音であったため，発話と誤検知されず精度の低下には至らなかったと考えられる．つまり，音声のみの手法では，環境音の大きさや種類などに精度が大きく影響されることがわかる．

4.2.3 画像と音声の情報を使用した推定

画像ベースの特徴量と音声ベースの特徴量両方を使用し,「環境音無」のテストデータの推定を行った結果を 10,「環境音無/発話無含」のテストデータの推定を行った結果を表 11,「環境音有/発話無含」のテストデータの推定を行った結果を表 12 に示す．平均の値を見ると，音のみの場合以上の精度となっている．

表 11 では，表 10 のデータと比較して精度の低下は見ら

表 4 画像のみでの推定：環境音無

	accuracy	precision	recall	F 値
被験者 A	0.624	0.654	0.549	0.591
被験者 B	0.559	0.693	0.217	0.324
被験者 C	0.587	0.670	0.349	0.453
被験者 D	0.671	0.773	0.486	0.593
被験者 E	0.523	0.513	0.882	0.649
被験者 F	0.648	0.847	0.365	0.505
被験者 G	0.589	0.673	0.345	0.451
被験者 H	0.685	0.745	0.565	0.642
被験者 I	0.592	0.631	0.443	0.520
被験者 J	0.812	0.816	0.813	0.811
平均	0.629	0.702	0.502	0.554

表 5 画像のみでの推定：環境音無/発話無含

	accuracy	precision	recall	F 値
被験者 A	0.476	0.458	0.261	0.327
被験者 B	0.473	0.385	0.096	0.151
被験者 C	0.426	0.271	0.098	0.141
被験者 D	0.445	0.389	0.193	0.257
被験者 E	0.534	0.519	0.930	0.666
被験者 F	0.571	0.682	0.266	0.379
被験者 G	0.569	0.642	0.310	0.413
被験者 H	0.439	0.398	0.241	0.300
被験者 I	0.597	0.631	0.466	0.536
被験者 J	0.822	0.798	0.867	0.829
平均	0.535	0.517	0.373	0.400

表 6 画像のみでの推定：環境音有/発話無含

	accuracy	precision	recall	F 値
被験者 A	0.609	0.632	0.536	0.574
被験者 B	0.611	0.727	0.363	0.472
被験者 C	0.727	0.709	0.779	0.740
被験者 D	0.671	0.752	0.515	0.602
被験者 E	0.557	0.533	0.932	0.678
被験者 F	0.648	0.742	0.456	0.562
被験者 G	0.632	0.715	0.439	0.540
被験者 H	0.770	0.754	0.802	0.776
被験者 I	0.624	0.622	0.636	0.628
被験者 J	0.827	0.775	0.925	0.842
平均	0.667	0.696	0.638	0.641

表 7 音声のみでの推定：環境音無

	accuracy	precision	recall	F 値
被験者 A	0.873	0.886	0.861	0.872
被験者 B	0.806	0.862	0.731	0.790
被験者 C	0.854	0.850	0.861	0.855
被験者 D	0.857	0.929	0.775	0.845
被験者 E	0.981	0.978	0.984	0.981
被験者 F	0.805	0.900	0.688	0.779
被験者 G	0.849	0.851	0.850	0.850
被験者 H	0.877	0.906	0.841	0.872
被験者 I	0.875	0.968	0.776	0.861
被験者 J	0.783	0.756	0.844	0.797
平均	0.856	0.889	0.821	0.850

表 8 音声のみでの推定：環境音無/発話無含

	accuracy	precision	recall	F 値
被験者 A	0.892	0.872	0.921	0.895
被験者 B	0.849	0.872	0.821	0.845
被験者 C	0.916	0.919	0.914	0.916
被験者 D	0.937	0.956	0.917	0.936
被験者 E	0.941	0.995	0.886	0.937
被験者 F	0.861	0.946	0.767	0.847
被験者 G	0.889	0.888	0.896	0.891
被験者 H	0.861	0.904	0.809	0.854
被験者 I	0.876	0.971	0.775	0.861
被験者 J	0.761	0.713	0.885	0.788
平均	0.878	0.903	0.859	0.877

表 9 音声のみでの推定：環境音有/発話無含

	accuracy	precision	recall	F 値
被験者 A	0.890	0.927	0.848	0.885
被験者 B	0.849	0.877	0.817	0.844
被験者 C	0.658	0.601	0.945	0.734
被験者 D	0.882	0.887	0.876	0.882
被験者 E	0.948	0.925	0.975	0.950
被験者 F	0.833	0.866	0.793	0.826
被験者 G	0.708	0.642	0.945	0.764
被験者 H	0.645	0.598	0.890	0.715
被験者 I	0.555	0.532	0.923	0.675
被験者 J	0.778	0.761	0.812	0.785
平均	0.775	0.762	0.882	0.806

れなかった．これは，音声情報を使用しているため，画像のみの推定の際にあった口を動かすが発話をしない部分に耐性ができたためだと考えられる．

表 12 は，音声情報のみでの「環境音有/発話無含」の推定結果である表 9 と比較すると精度が向上している．画像ベースの特徴量を加えることで環境音のノイズによる影響を抑えることができたと考えられる．

各被験者のデータを見た場合，被験者 I は環境音が存在する状況下での結果である表 9 と表 12 を比較しても精度の向上が見られなかった．被験者 I の環境音は，表 9 の結果を見ても，ノイズの影響がかなり大きいデータとなっている．今回の提案手法では，多くのデータにおいては環境音のノイズによる影響を抑えることができたが，被験者 I のような極端にノイズの影響を受けたデータに対しては効果がなかった．

4.3 追加実験

本研究の手法では，画像の特徴量としてオプティカルフローを用いたため，マスクの種類に依存せずに推定が行えると考えられる．追加実験としてデータセットのうち不織布マスクを着用した状態での動画を訓練データに，ウレタンマスクを着用した状態での動画をテストデータとした場

表 10 画像と音声での推定：環境音無

	accuracy	precision	recall	F 値
被験者 A	0.914	0.976	0.848	0.907
被験者 B	0.806	0.834	0.769	0.799
被験者 C	0.883	0.964	0.796	0.872
被験者 D	0.843	0.972	0.706	0.818
被験者 E	0.970	0.968	0.972	0.970
被験者 F	0.715	0.975	0.441	0.607
被験者 G	0.841	0.965	0.707	0.816
被験者 H	0.893	0.933	0.848	0.888
被験者 I	0.909	0.956	0.859	0.905
被験者 J	0.895	0.838	0.984	0.904
平均	0.867	0.938	0.793	0.849

表 11 画像と音声での推定：環境音無/発話無含

	accuracy	precision	recall	F 値
被験者 A	0.927	0.986	0.867	0.923
被験者 B	0.832	0.811	0.871	0.839
被験者 C	0.911	0.949	0.870	0.908
被験者 D	0.906	0.952	0.855	0.901
被験者 E	0.927	0.977	0.874	0.923
被験者 F	0.717	0.994	0.437	0.606
被験者 G	0.860	0.971	0.744	0.842
被験者 H	0.855	0.908	0.793	0.846
被験者 I	0.897	0.957	0.833	0.890
被験者 J	0.870	0.808	0.976	0.883
平均	0.870	0.931	0.812	0.856

表 12 画像と音声での推定：環境音有/発話無含

	accuracy	precision	recall	F 値
被験者 A	0.923	0.983	0.860	0.918
被験者 B	0.914	0.979	0.846	0.908
被験者 C	0.756	0.685	0.948	0.795
被験者 D	0.850	0.983	0.713	0.826
被験者 E	0.959	0.932	0.992	0.961
被験者 F	0.801	0.994	0.605	0.752
被験者 G	0.864	0.840	0.906	0.870
被験者 H	0.864	0.883	0.841	0.861
被験者 I	0.502	0.501	1.000	0.668
被験者 J	0.850	0.778	0.985	0.869
平均	0.828	0.856	0.870	0.843

合での実験を行った．表 13 に訓練データ数とテストデータ数を示す．

追加実験を行った結果を表 14 に示す．画像のみの推定結果について，表 4 の不織布の結果と比較すると精度が低い傾向が見られる．これは，ウレタンマスクは訓練データに含まれていないのが原因と考えられる．

しかし，訓練データに不織布マスクのものしか含まれてない場合においても，音声のみの情報を使用した発話推定よりも，画像と音声両方の特徴量を使用した場合の方が高い精度で推定が行われていることが確認できる．被験者が少なく個人差による影響の確認が必要ではあるが，マスクの種類に依存せず，画像の情報を追加することで精度の向上できると考えられる．

表 13 追加実験の訓練データ数とテストデータ数

訓練データ	環境音無し (件)	環境音無し 発話無し含	環境音あり 発話無し含む
11264	2396	672	682

表 14 ウレタンマスク着用時の発話推定結果

環境音無				
特徴量	正解率	precision	recall	F 値
画像	0.497	0.490	0.234	0.314
音声	0.788	0.751	0.869	0.805
画像＋音声	0.825	0.905	0.728	0.806
環境音無/発話無含				
特徴量	正解率	precision	recall	F 値
画像	0.523	0.544	0.268	0.357
音声	0.814	0.785	0.868	0.824
画像＋音声	0.856	0.945	0.757	0.840
環境音有/発話無含				
特徴量	正解率	precision	recall	F 値
画像	0.556	0.601	0.333	0.424
音声	0.833	0.854	0.804	0.828
画像＋音声	0.915	0.967	0.860	0.910

5. まとめ

本研究では，マスク着用環境下において画像と音声の情報を用いた発話推定を行った．画像の情報として，本研究ではオプティカルフローを使用し，マスクのオプティカルフローを特徴量として推定を行った．音声情報としては，指定した範囲の周波数の音の大きさを特徴量として推定を行った．また，本研究では精度比較のために画像の情報のみを用いた推定と音声情報のみを用いた推定も行った．

画像の情報のみでの推定では良い精度の推定が行えなかった．音声情報のみでの推定では，環境音が無い場合は高い精度で推定が行えたが，環境音がある場合は環境音がノイズとなり，精度が低下した．画像と音声情報を組み合わせた推定では，環境音がある場合においても高い精度で推定が行えた．

今後の課題としては，画像のみでの推定の精度が低いため，オプティカルフローの取得法の改良や，オプティカルフローではない特徴量の場合の調査が必要である．本研究では音声情報として取得する周波数の範囲を指定したが，A-VAD で使用されている手法を用いて音声情報による推定の精度向上させた場合における，画像の情報を加えた推定の有効性の確認も必要である．また，追加実験で本研究の手法の場合，マスクの種類に依存せずに発話推定が行える可能性が確認できたので，その調査を行っていきたい．

謝辞

本研究は科研費 23K11368 の一部です.

参考文献

[1] Meduri, S. S. and Ananth, R.: A Survey and Evaluation of Voice Activity Detection Algorithms, (online), available from ⟨https://api.semanticscholar.org/CorpusID:102345548⟩ (2012).
[2] Ali, Z. and Talha, M.: Innovative method for unsupervised voice activity detection and classification of audio segments, *Ieee Access*, Vol. 6, pp. 15494–15504 (2018).
[3] Alimi, S. and Awodele, O.: Voice Activity Detection: Fusion of Time and Frequency Domain Features with A SVM Classifier, *Comput. Eng. Intell. Syst*, Vol. 13, No. 3, pp. 20–29 (2022).
[4] Gudepu, P. R., Koroth, J. M., Sabu, K. and Shaik, M. A. B.: Dynamic Encoder RNN for Online Voice Activity Detection in Adverse Noise Conditions, *Proc. INTERSPEECH 2023*, pp. 5052–5056 (online), DOI: 10.21437/Interspeech.2023-2466 (2023).
[5] Zhou, H., Du, J., Chen, H., Jing, Z., Xiong, S. and Lee, C.-H.: Audio-Visual Information Fusion Using Cross-Modal Teacher-Student Learning for Voice Activity Detection in Realistic Environments, *Proc. Interspeech 2021*, pp. 341–345 (online), DOI: 10.21437/Interspeech.2021-592 (2021).
[6] Tao, F. and Busso, C.: Bimodal Recurrent Neural Network for Audiovisual Voice Activity Detection, *Proc. Interspeech 2017*, pp. 1938–1942 (online), DOI: 10.21437/Interspeech.2017-1573 (2017).
[7] 元吉大介，嶋田和孝，榎田修一，江島俊朗，遠藤　勉：ロボットとの対話のための発話推定に関する事例研究，画像の認識・理解シンポジウム (MIRU2008)，pp. 1015–1020 (2008).
[8] 二宮芳樹，坂　義秀，前野俊希，根木大輔，宮島千代美，森　健策，北坂孝幸，末永康仁：音声と画像の統合によるドライバの発話区間検出，映像情報メディア学会誌，Vol. 62, No. 3, pp. 435–441 (2008).
[9] Lucas, B. D. and Kanade, T.: An iterative image registration technique with an application to stereo vision, *IJCAI'81: 7th international joint conference on Artificial intelligence*, Vol. 2, pp. 674–679 (1981).
[10] Farnebäck, G.: Two-Frame Motion Estimation Based on Polynomial Expansion, *Image Analysis* (Bigun, J. and Gustavsson, T., eds.), Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 363–370 (2003).
[11] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780 (online), DOI: 10.1162/neco.1997.9.8.1735 (1997).